university of groningen

University Medical Center Groningen

University of Groningen

Data Validation Beyond Big Data

Valentijn, Edwin A.

Published in:
VST in the Era of the Large Sky Surveys

DOI:
10.5281/zenodo.1303323

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2018

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):
Valentijn, E. A. (2018). Data Validation Beyond Big Data. In *VST in the Era of the Large Sky Surveys: Proceedings of the conference held 5-8 June, 2018 in Naples, Italy* (pp. 17)
https://doi.org/10.5281/zenodo.1303323

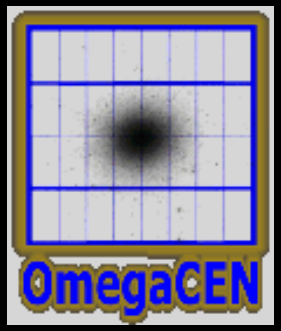
Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license. More information can be found on the University of Groningen website: https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment.

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Download date: 05-05-2023



Data validation beyond Big Data

Edwin A. Valentijn

Kapteyn Astronomical Institute

6 June 2018 VST in the era of large sky surveys- Napoli

STORY LINES

- processing/archiving/distribution:

  - AstroWISE- KiDs - Ou-Ext – Euclid

- data validation:

  - lineage - OU-Ext - Euclid- Facts and Fakes

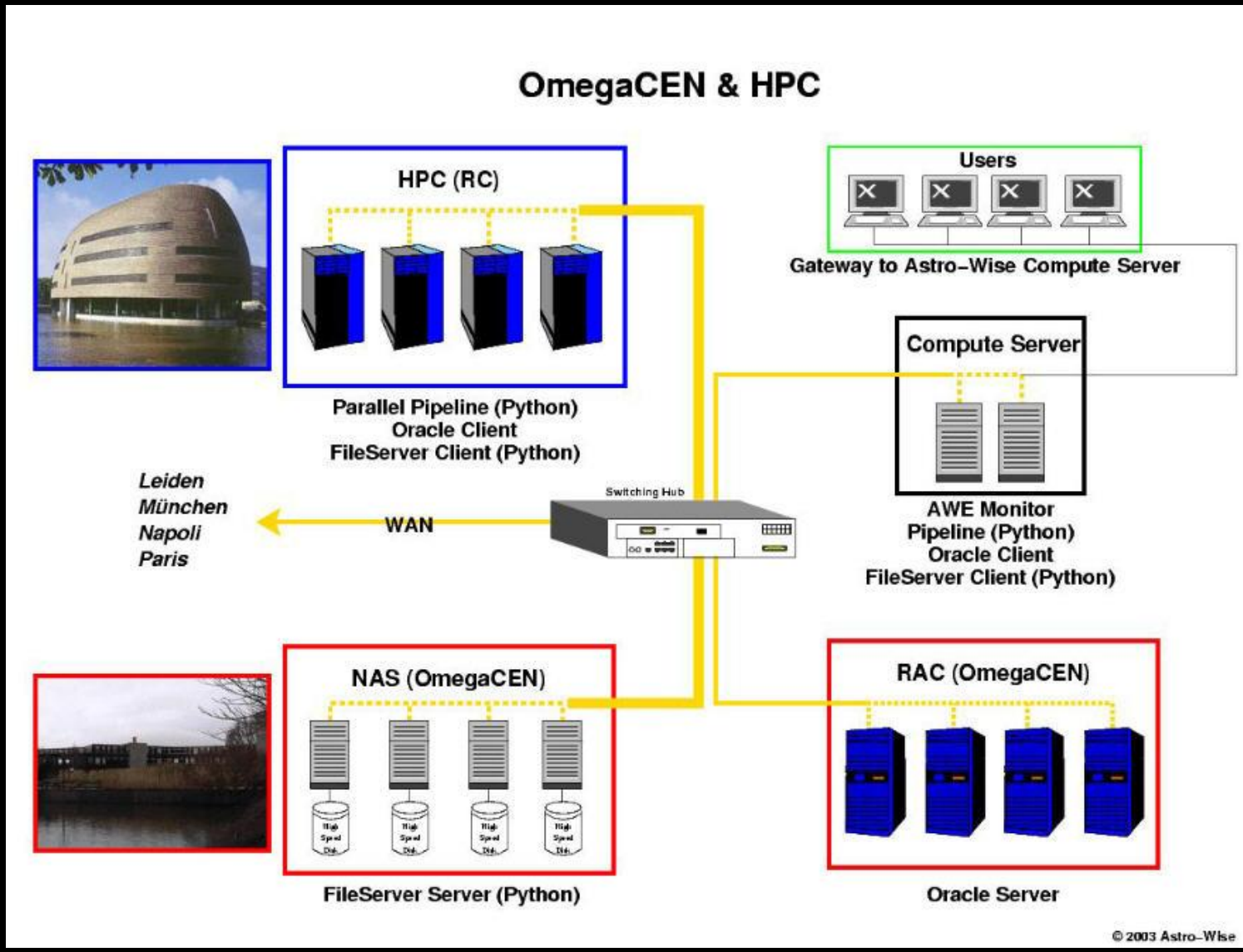
Sequence of hypes:

GRID - Big Data - Machine learning -> data validation

The Datacentric approach

local networks and distributed

2003
RUG-CIT



Astro-WISE – Data federations

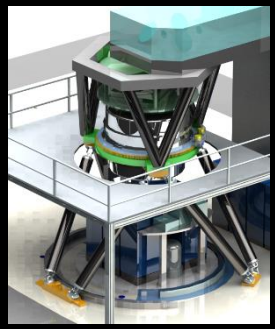
Distributed Information Systems - handling surveys

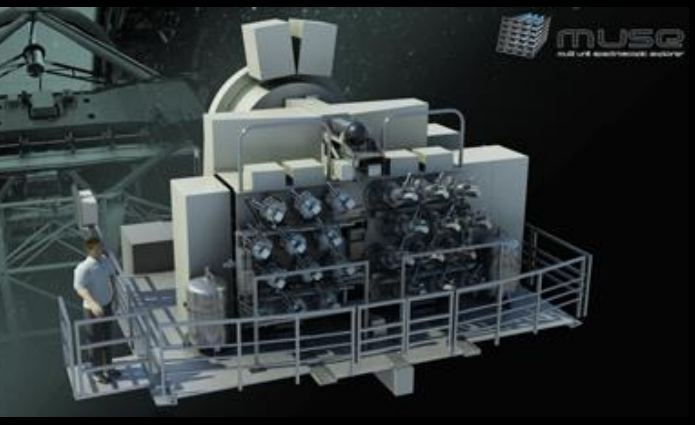
since 2003 - it works

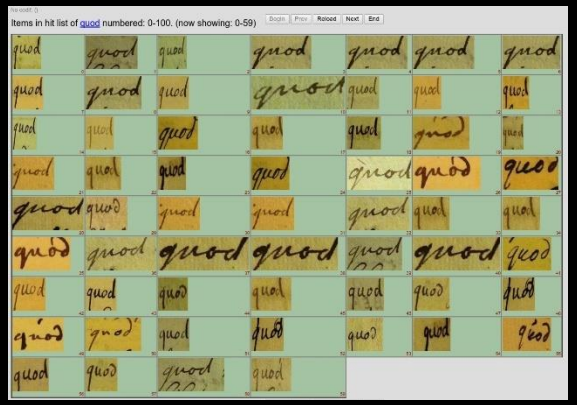
OmegaCEN@Kapteyn datacenter ~15-20 fte

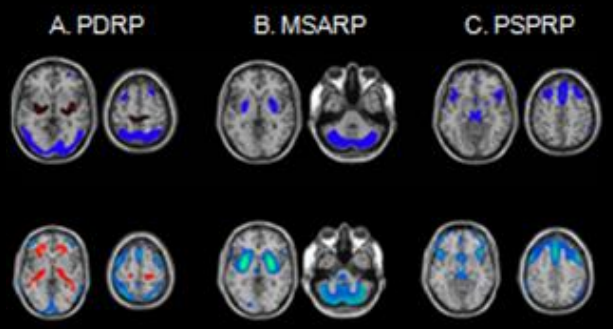
KiDS - ESO – OmegaCAM@VST
MUSE - ESO - VLT
Lofar - LTA - Astron
Glimps - AI Handwritten text – Lifelines DNA
Target Holding

-> Euclid - ESA
-> Micado - ESO - ELT





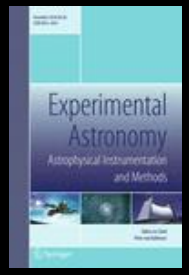




all published

http://www.astro-wise.org Manuals & tutorials

http://www.rug.nl/target Target Consortium



Experimental Astronomy - Vol. 35, 2013

All papers are online

Astroinformatics
Proceedings IAU Symposium No. 325, 2016
M. Brescia, S.G. Djorgovski, E. Feigelson, G. Longo & S. Cavuoti, eds.
© International Astronomical Union 2017
doi:10.1017/S1743921317000254

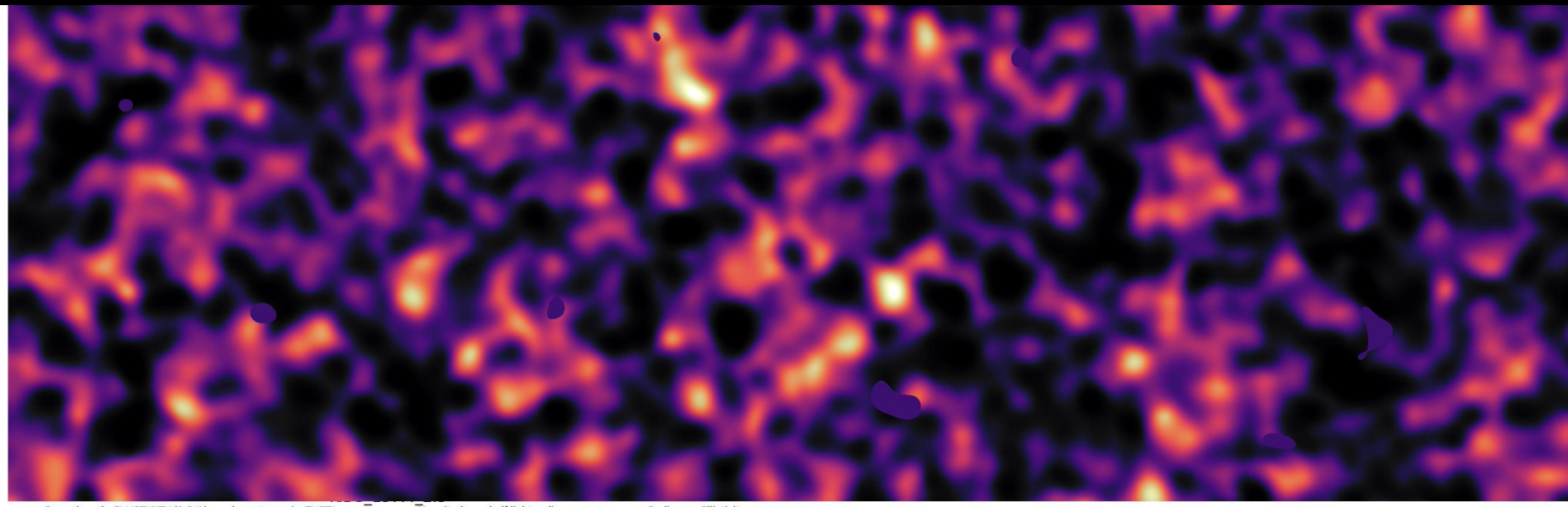
Target and (Astro-)WISE technologies Data federations and its applications

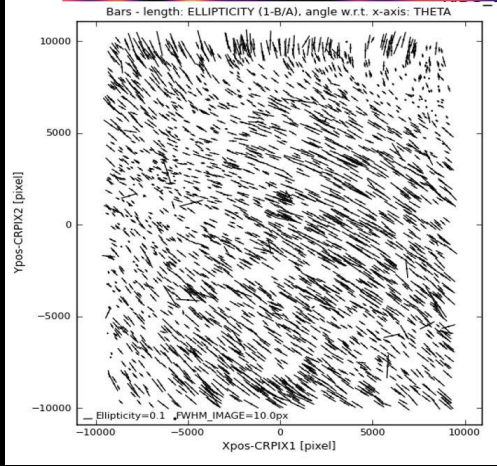
E. A. Valentijn[1], K. Begeman[1], A. Belikov[1], D. R. Boxhoorn[1], J. Brinchmann[2], J. McFarland[1], H. Holties[3], K. H. Kuijken[2], G. Verdoes Kleijn[1], W-J. Vriend[1], O. R. Williams[4], J. B. T. M. Roerdink[5], L. R. B. Schomaker[6], M. A. Swertz[7], A. Tsyganov[4] and G. J. W. van Dijk[8]

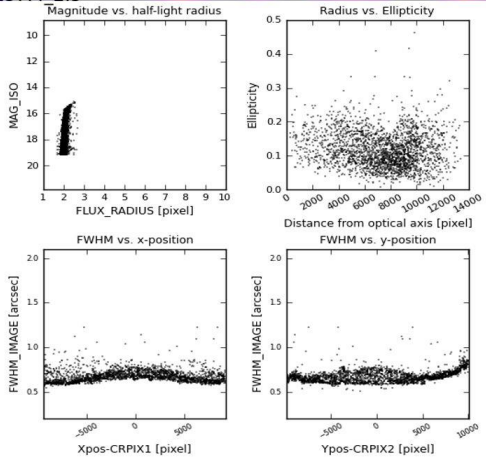
[1]Kapteyn Astronomical Institute, University of Groningen, email: valentyn@astro.rug.nl
[2]Leiden Observatory, Leiden University
[3]ASTRON, Dwingeloo
[4]Center for Information Technology, University of Groningen
[5]Johann Bernoulli Institute, University of Groningen
[6]ALICE, University of Groningen
[7]University Medical Center Groningen, University of Groningen
[8]Target Holding, Groningen

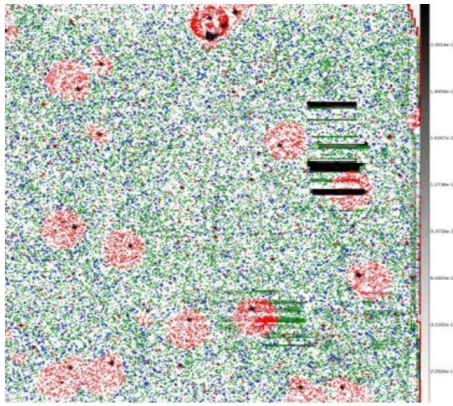
Astroinformatics 2016
IAU symposium 325
Datafederations
Valentijn et al. 2017

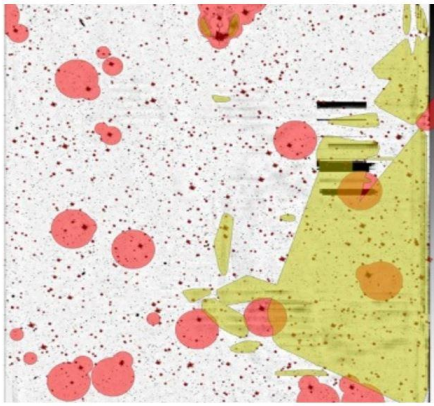
KiDS Quality control DR1-DR2-DR3 OmegaCAM@VST 740 sq deg











Links as workhorse in data federations

The Universe as a spreadsheet

ERCIM News 2006

AstroWISE *Chaining to the Universe*

ADASS XVI ASP Conference Series,

15-18 October 2006 in Tucson, Arizona, USA.

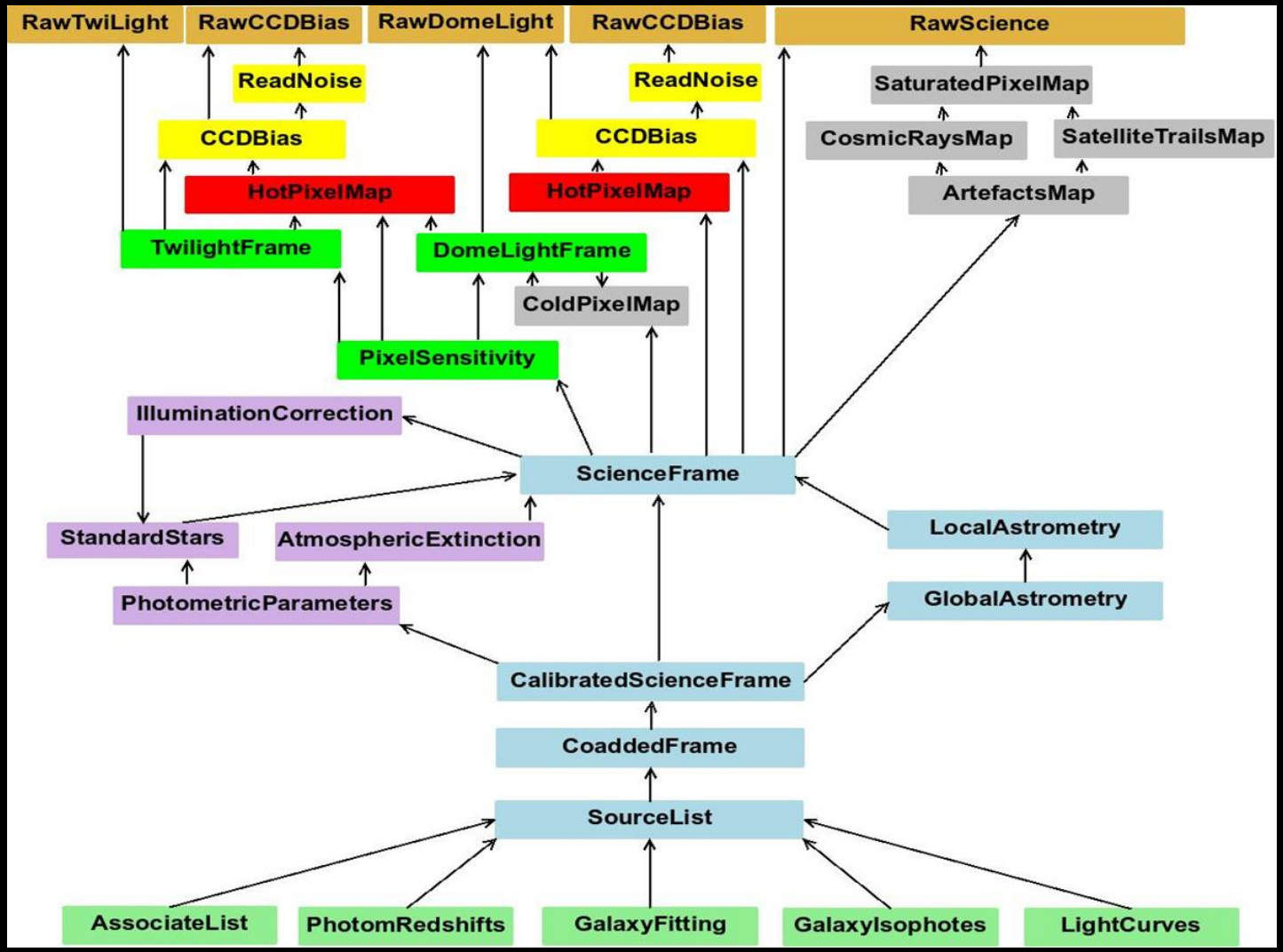
- Distributed Information Systems
  - Users, computers, storage
- Processing and Quality control
- Reproducable (re-processing)

2018: Open Science - **FAIR** principles

Findable **A**ccessable **I**nteroperable **R**eproducable

The universe as a spreadsheet

Target Diagram/Data lineage /backward chaining ++ programming - dependencies

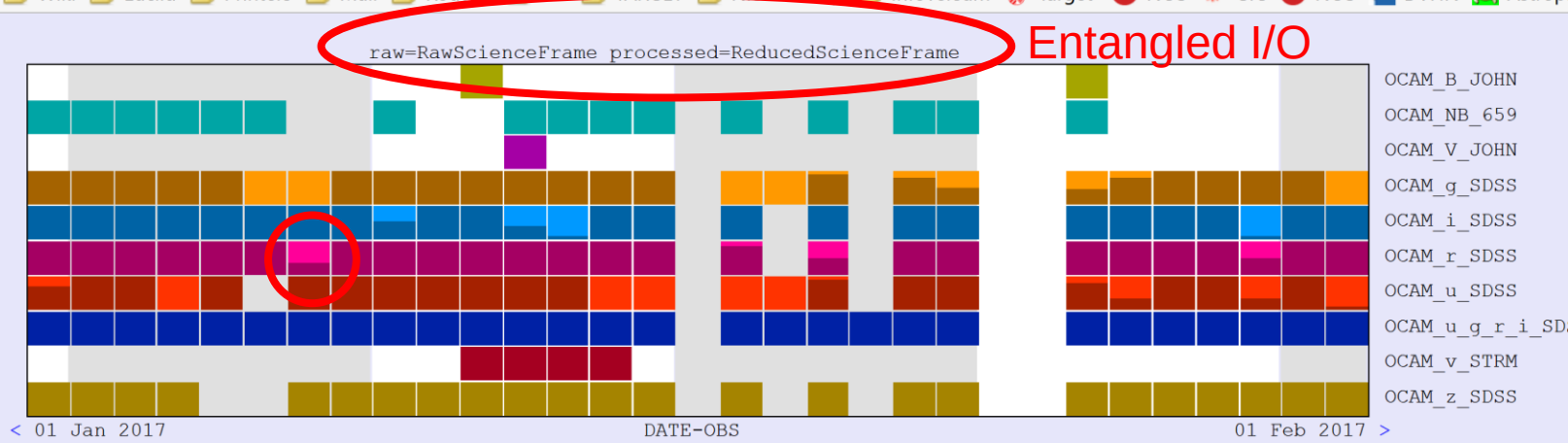


QUERY / INFORMATION

PROCESSING

Beheer uw | Google Ma | University | Eklat an de | Programme | my citizenl | NOS.nl - Ni | [1710.0590 | DESHIMA z | Facebook | New e
process.astro-wise.org/Process
Euclid conference Bonn
OCam | OCen | EV | Wiki | Euclid | Printers | Mail | Routers | AA | TARGET | Nieuwe map | Infoversum | Target | NOS | GIC | NOS | DVHN | Astrop

Entangled I/O



Specify Target

Specify a period and click show. For the selected period all available observations will be shown in the above view. Each block corresponds to one or a set of observations with a specific filter or observing block. Click on a block to get an overview of the possible targets. You can also use the extended query form.

raw	processed	
192	0	OCAM_B_JOHN JohnsonB
9184	0	OCAM_NB_659 UnknownNB659
32	0	OCAM_V_JOHN JohnsonV
6624	2400	OCAM_g_SDSS SloanG
10624	2048	OCAM_i_SDSS SloanI
11008	640	OCAM_r_SDSS SloanR
7808	2595	OCAM_u_SDSS SloanU
2976	0	OCAM_u_g_r_i_SDSS SloanUGR
128	0	OCAM_v_STRM StromgrenV
1376	0	OCAM_z_SDSS SloanZ

Period Selection (DATE-OBS)

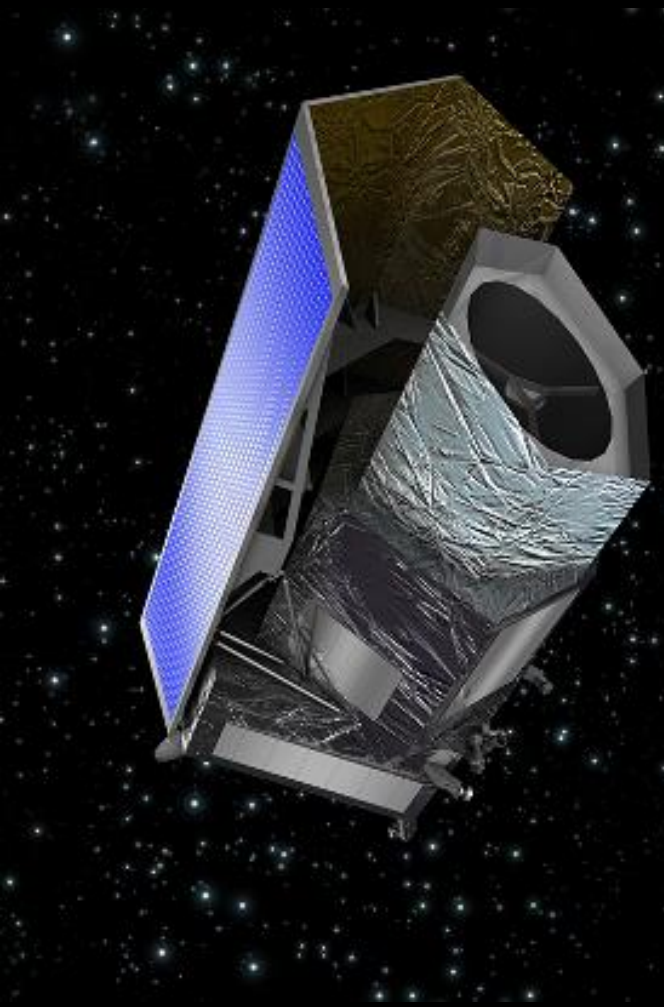
Year	Quarter	Month	Week
2017	<none>	1 jan	<none>

Optional Settings

Name	Value
Filter	<none>
Group by	(•) Filter ( ) Observing Block ( ) Template
Filtering	[x] Flagged data [ ] Project only

Show

Euclid

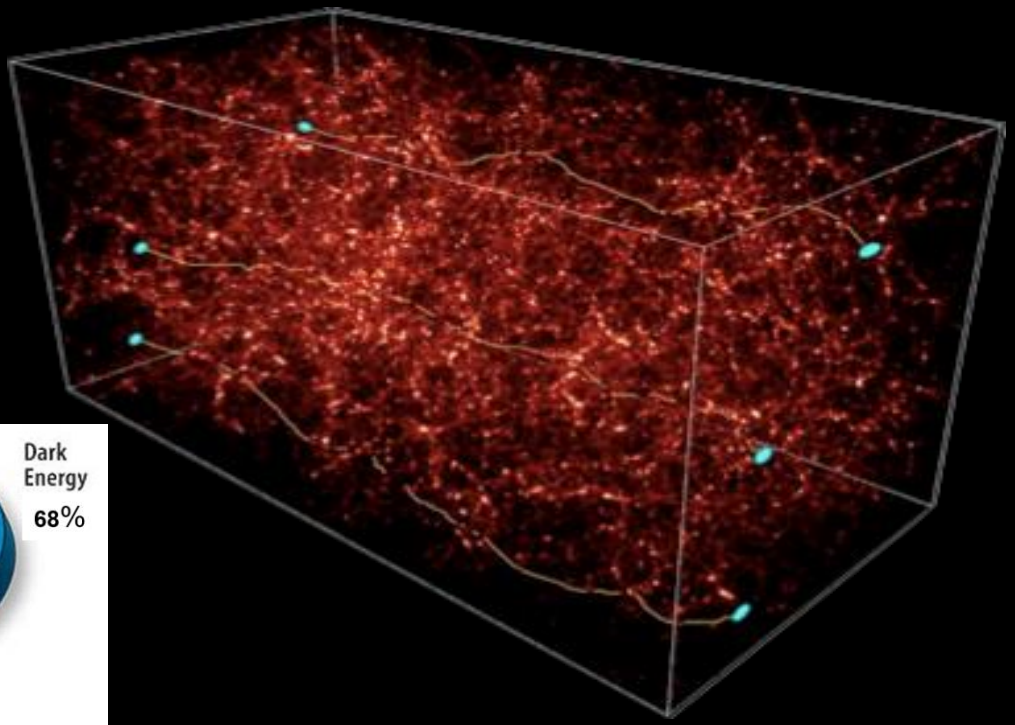


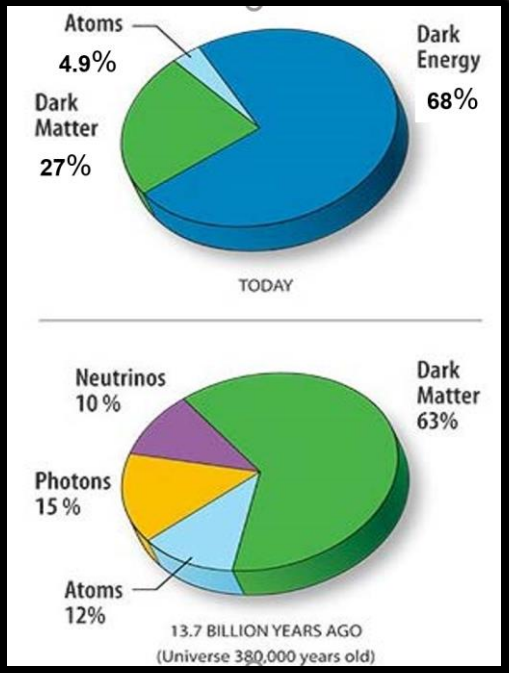
ESA launch in May 2021

Euclid Archive System (EAS)

- data centric information system
- many of the WISE concepts
- prototype uses Astro-WISE
- db hosted in the Euclid SDC-NL in Groningen

Weak gravitational lensing as probe of dark matter





KiDS: < 100 10^6 redshifts

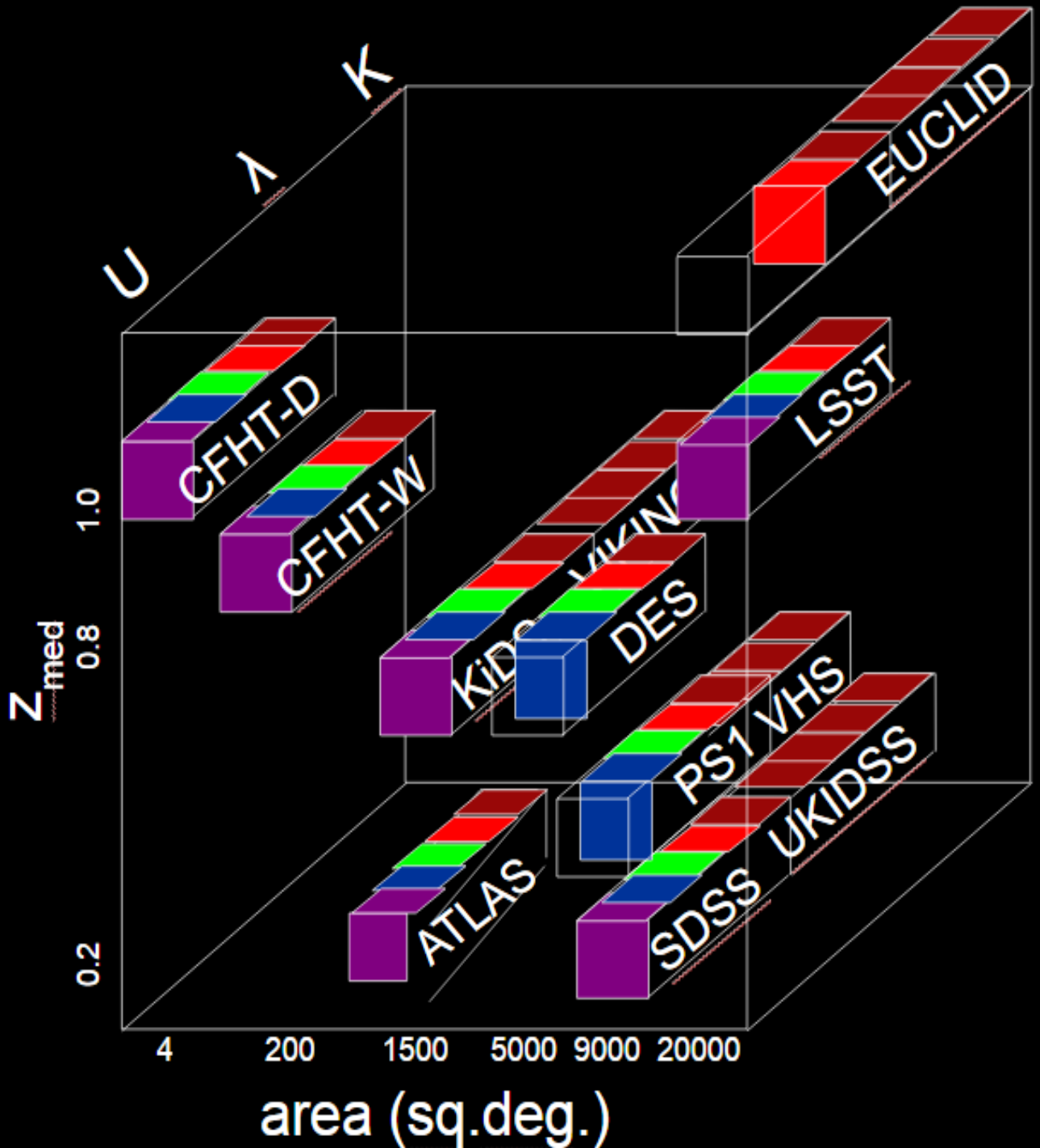
EUCLID: 1.5 10^9 redshifts - phot- z

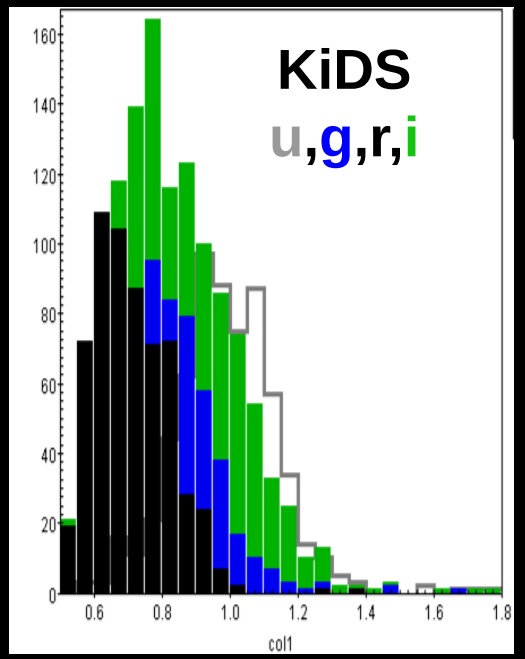
Ground based data – OU-Ext

Every galaxy has its own 4 PSFs

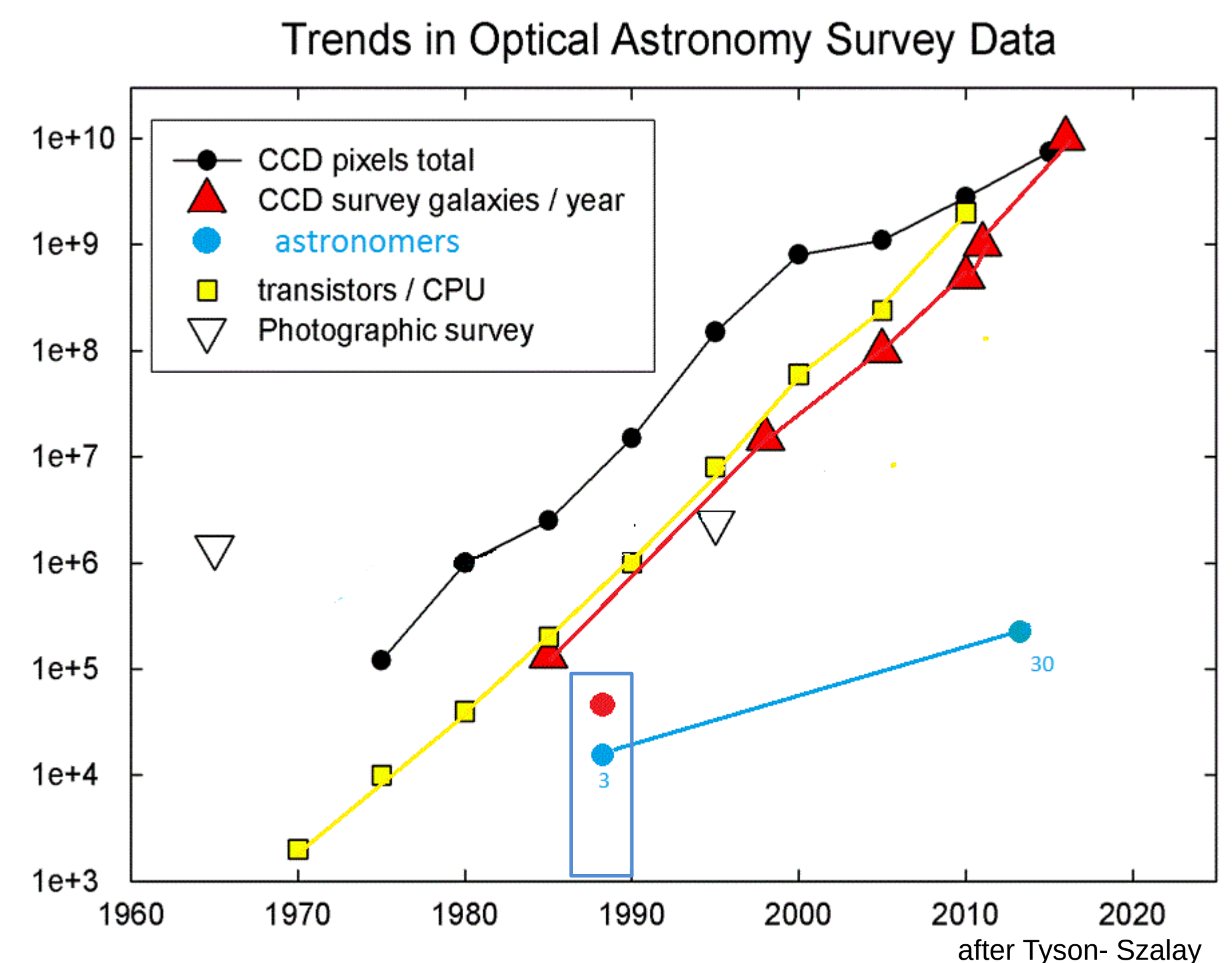
QC- bias – re-processing

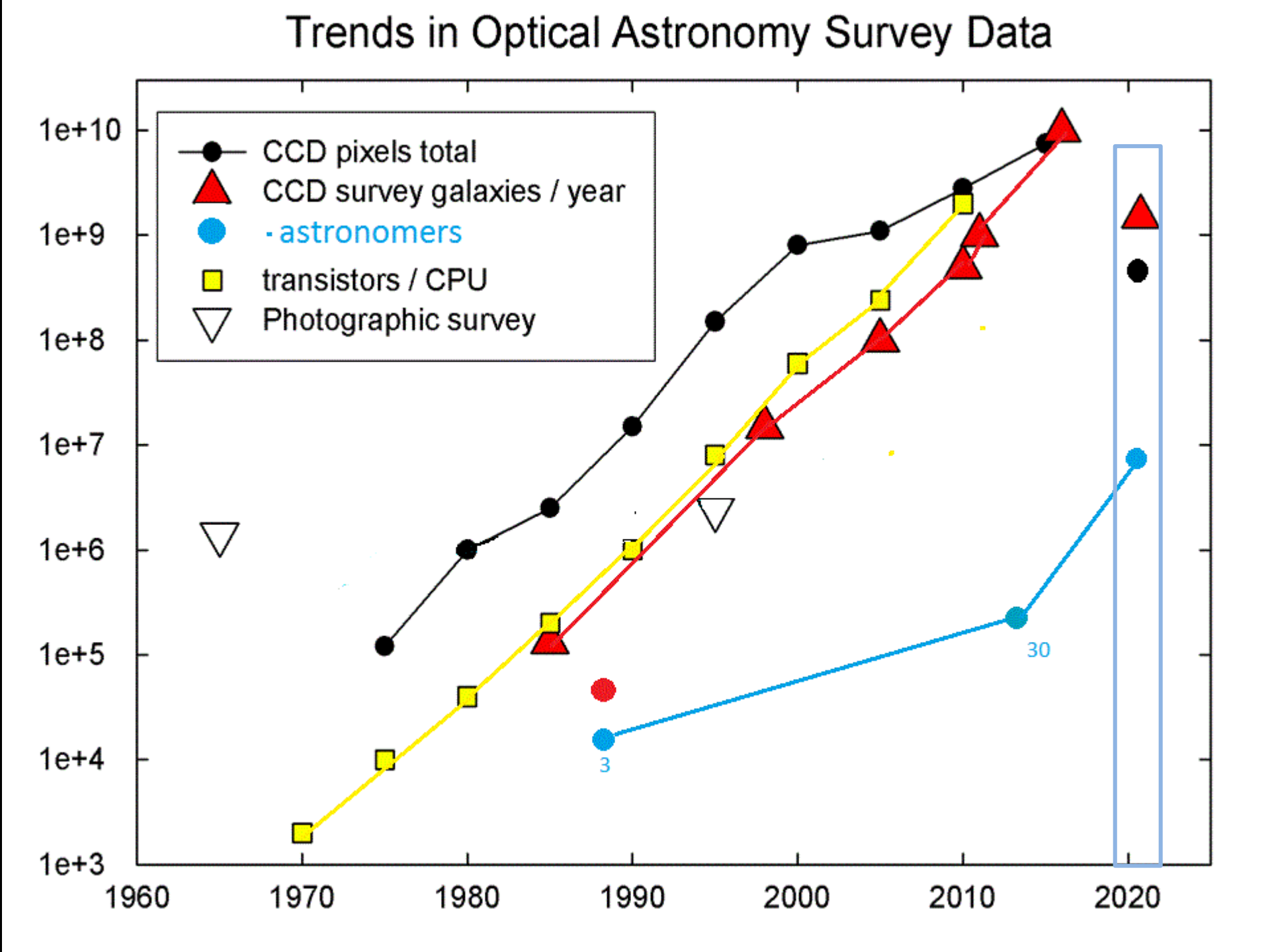
KiDS/VIKING



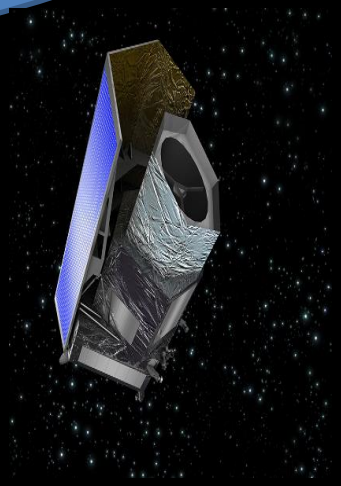


Seeing (“)


Trends in Optical Astronomy Survey Data
CCD pixels total
CCD survey galaxies / year
astronomers
transistors / CPU
Photographic survey
1e+10
1e+9
1e+8
1e+7
1e+6
1e+5
1e+4
1e+3
1960
1970
1980
1990
2000
2010
2020
30
3
after Tyson- Szalay


Trends in Optical Astronomy Survey Data
CCD pixels total
CCD survey galaxies / year
astronomers
transistors / CPU
Photographic survey
1e+10
1e+9
1e+8
1e+7
1e+6
1e+5
1e+4
1e+3
1960
1970
1980
1990
2000
2010
2020
3
30

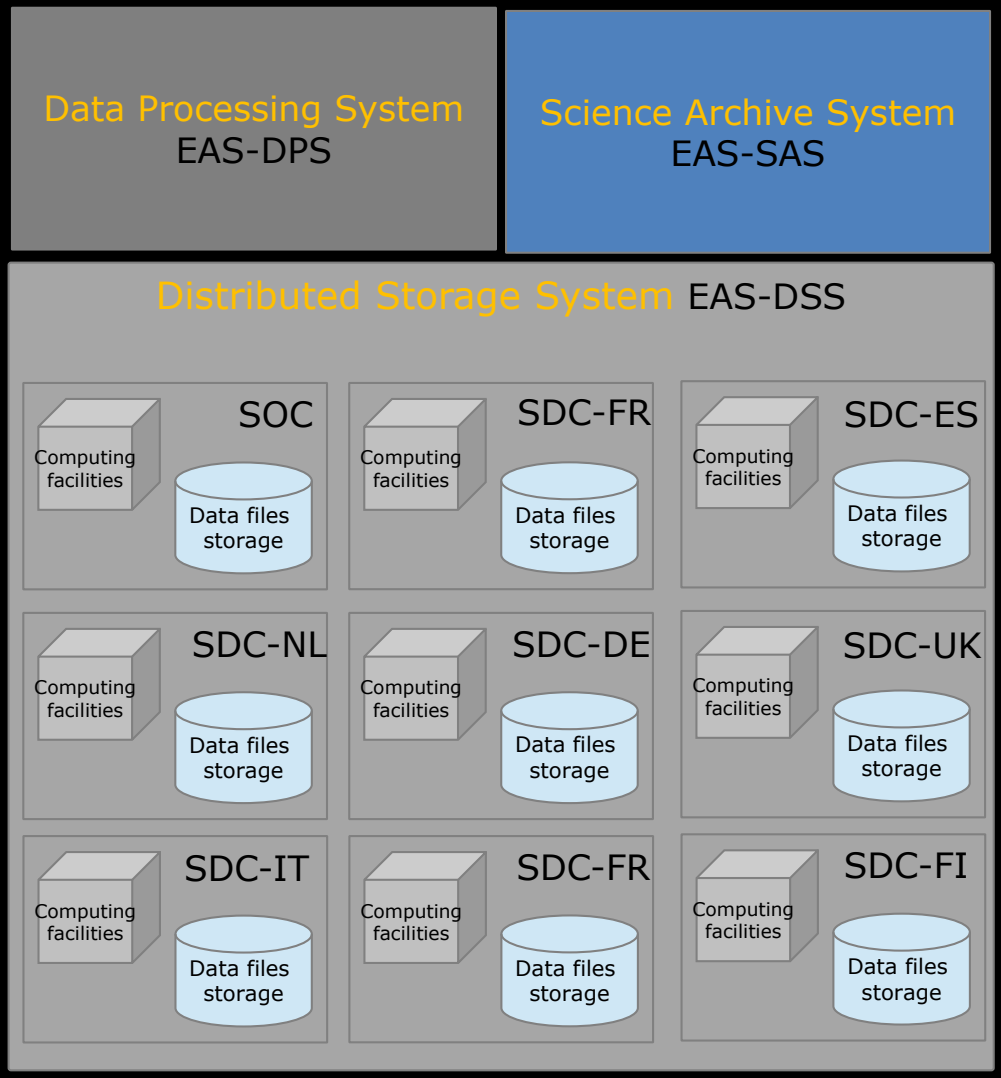
Distributed communities acces-proces-calibrate-analyse publish

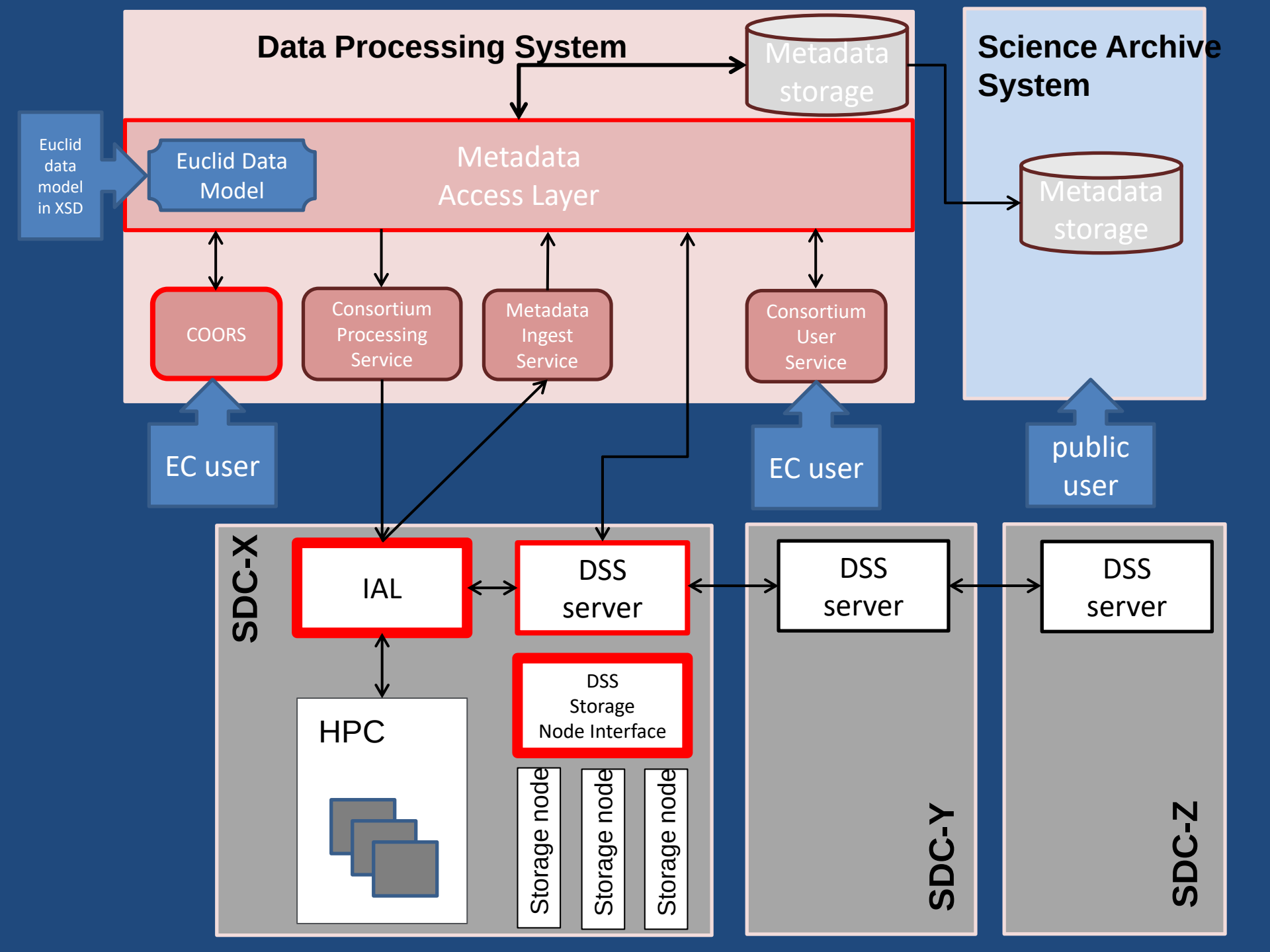


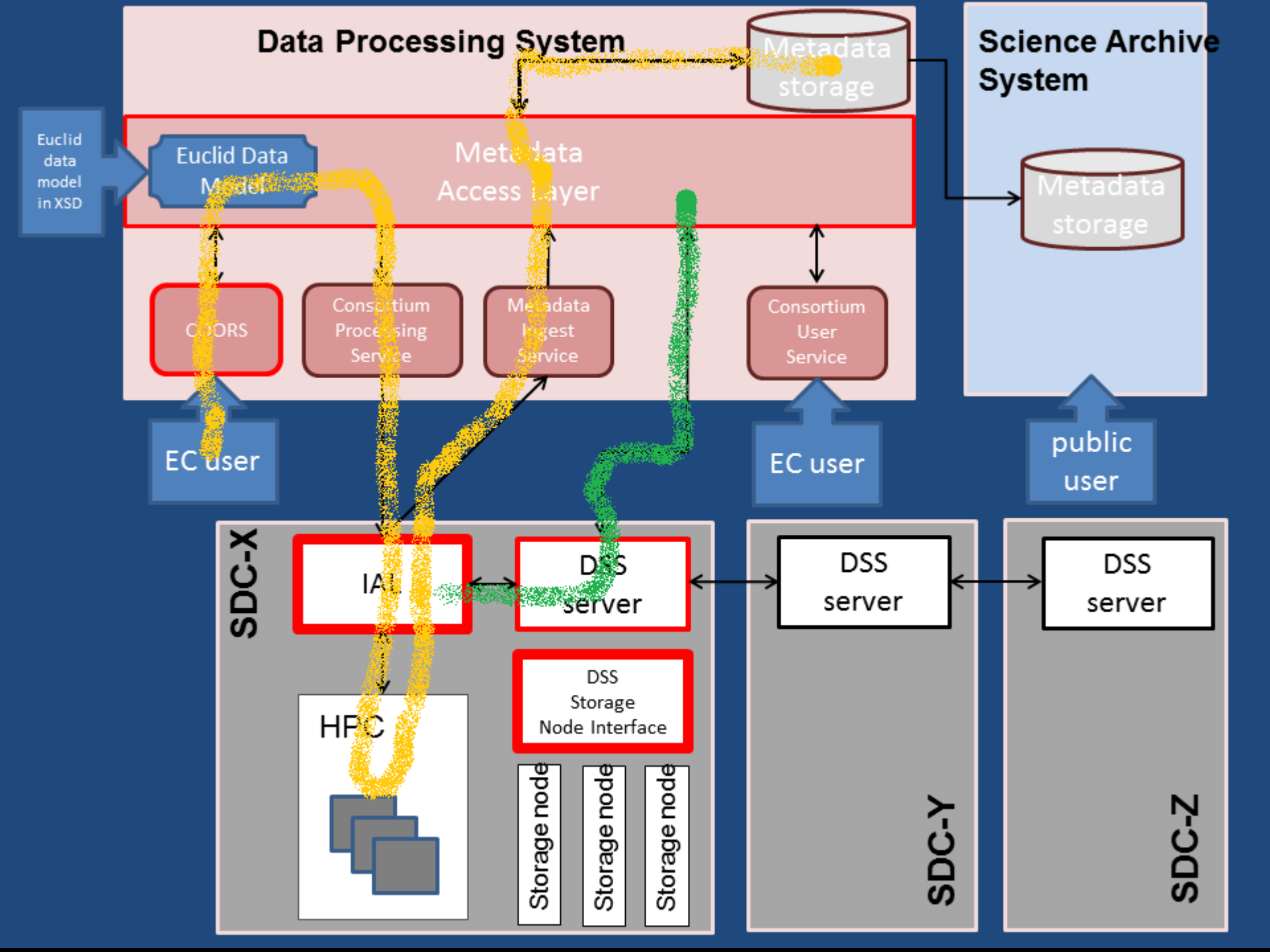
Euclid:

- 1500 registered members and growing
- 200 laboratories/departments
- 16 countries contributing
- NASA/US: provides the IR detectors.

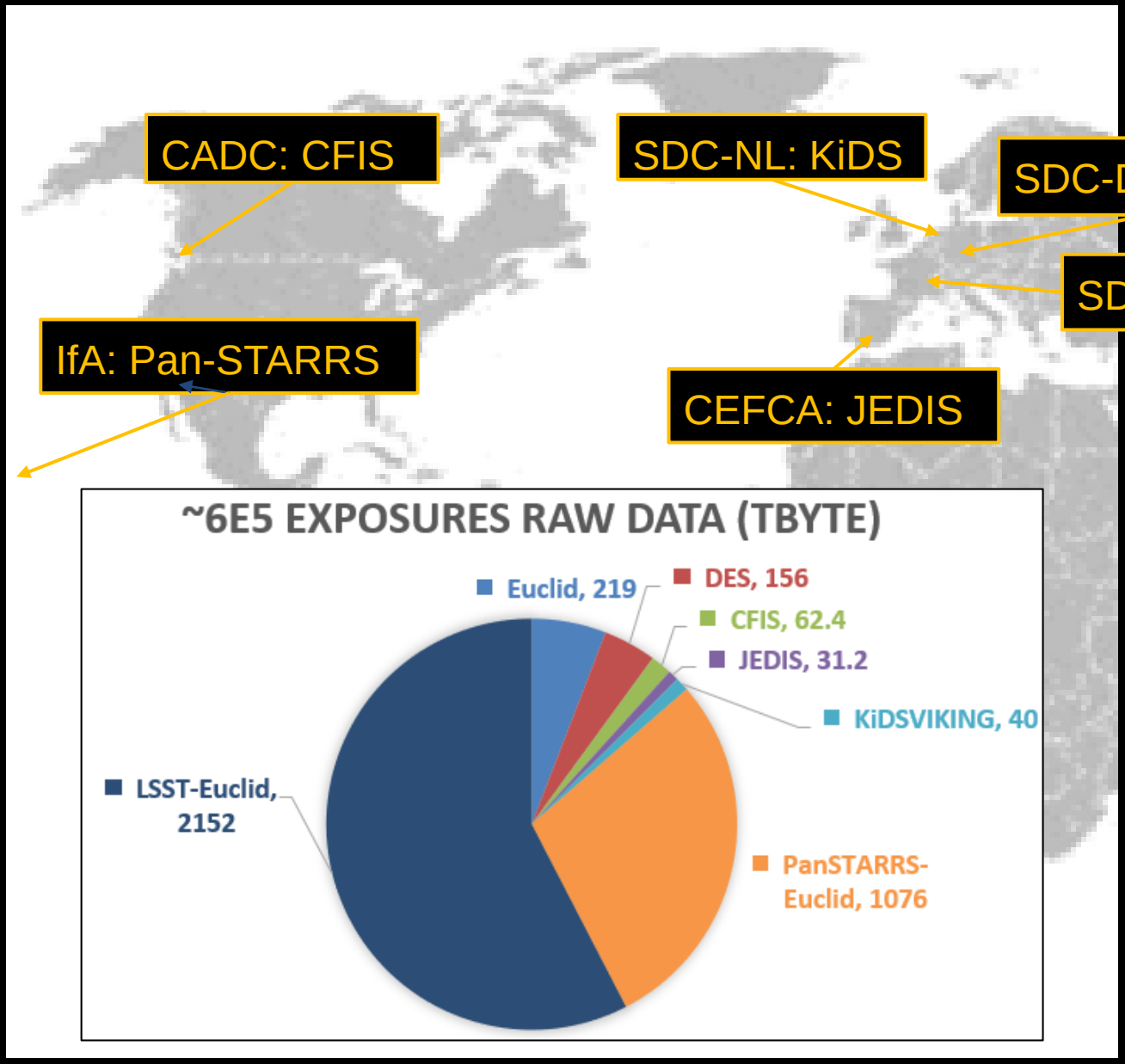
Euclid Archive system – EAS – lay out



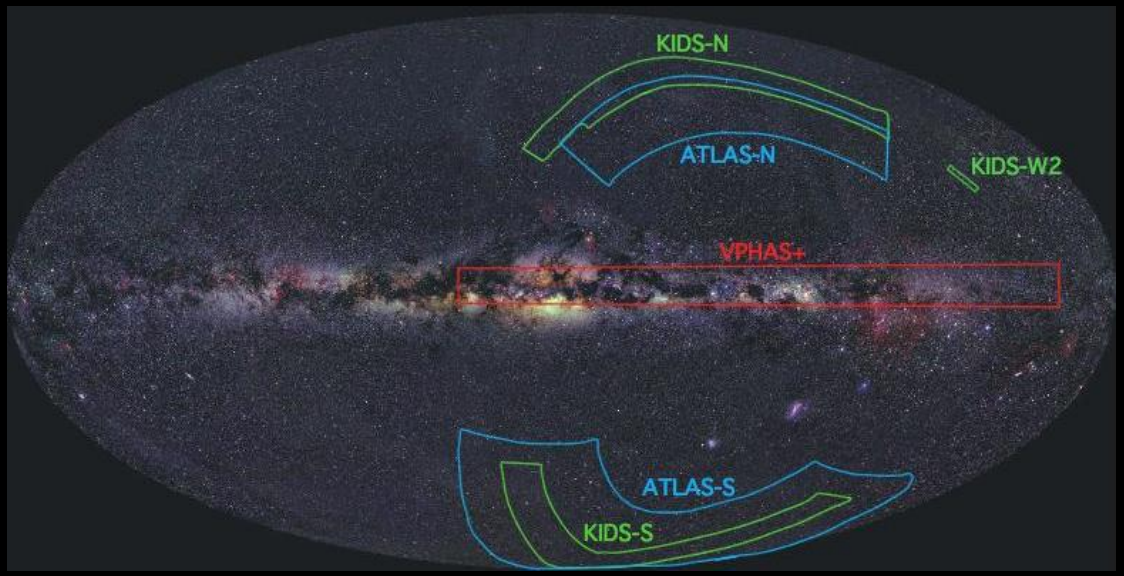

Data Processing System
Metadata storage
Science Archive System
Euclid data model in XSD
Euclid Data Model
Metadata Access Layer
Metadata storage
COORS
Consortium Processing Service
Metadata Ingest Service
Consortium User Service
EC user
EC user
public user
SDC-X
IAL
DSS server
DSS server
DSS server
HPC
DSS Storage Node Interface
Storage node
Storage node
Storage node
SDC-Y
SDC-Z

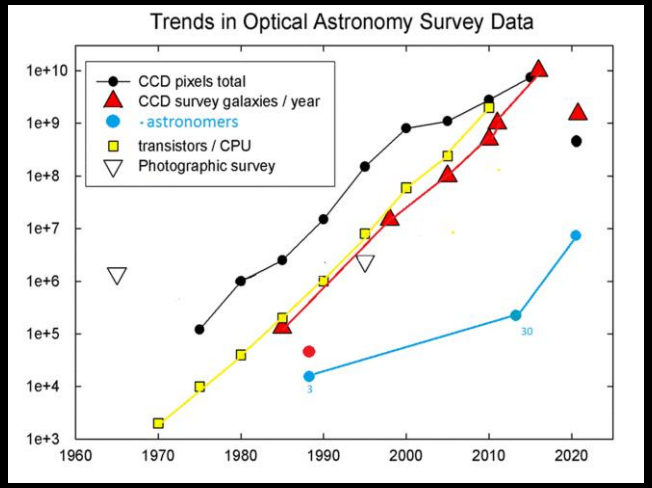

Data Processing System
Metadata storage
Science Archive System
Euclid data model in XSD
Euclid Data Model
Metadata Access Layer
Metadata storage
CODORS
Consortium Processing Service
Metadata Ingest Service
Consortium User Service
EC user
EC user
public user
SDC-X
IAL
DSS server
DSS server
DSS server
DSS Storage Node Interface
HPC
Storage node
Storage node
Storage node
SDC-Y
SDC-Z

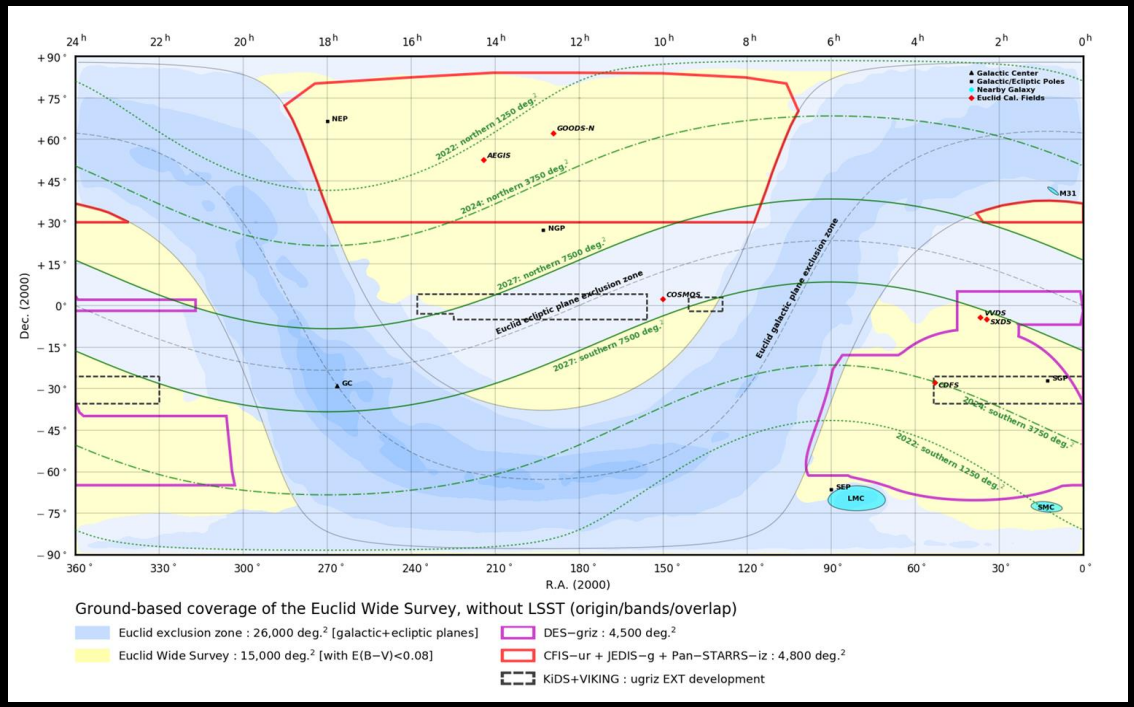
Euclid-EXT: massive pixel volumes - distributed archives



From KiDS to Euclid-EXT

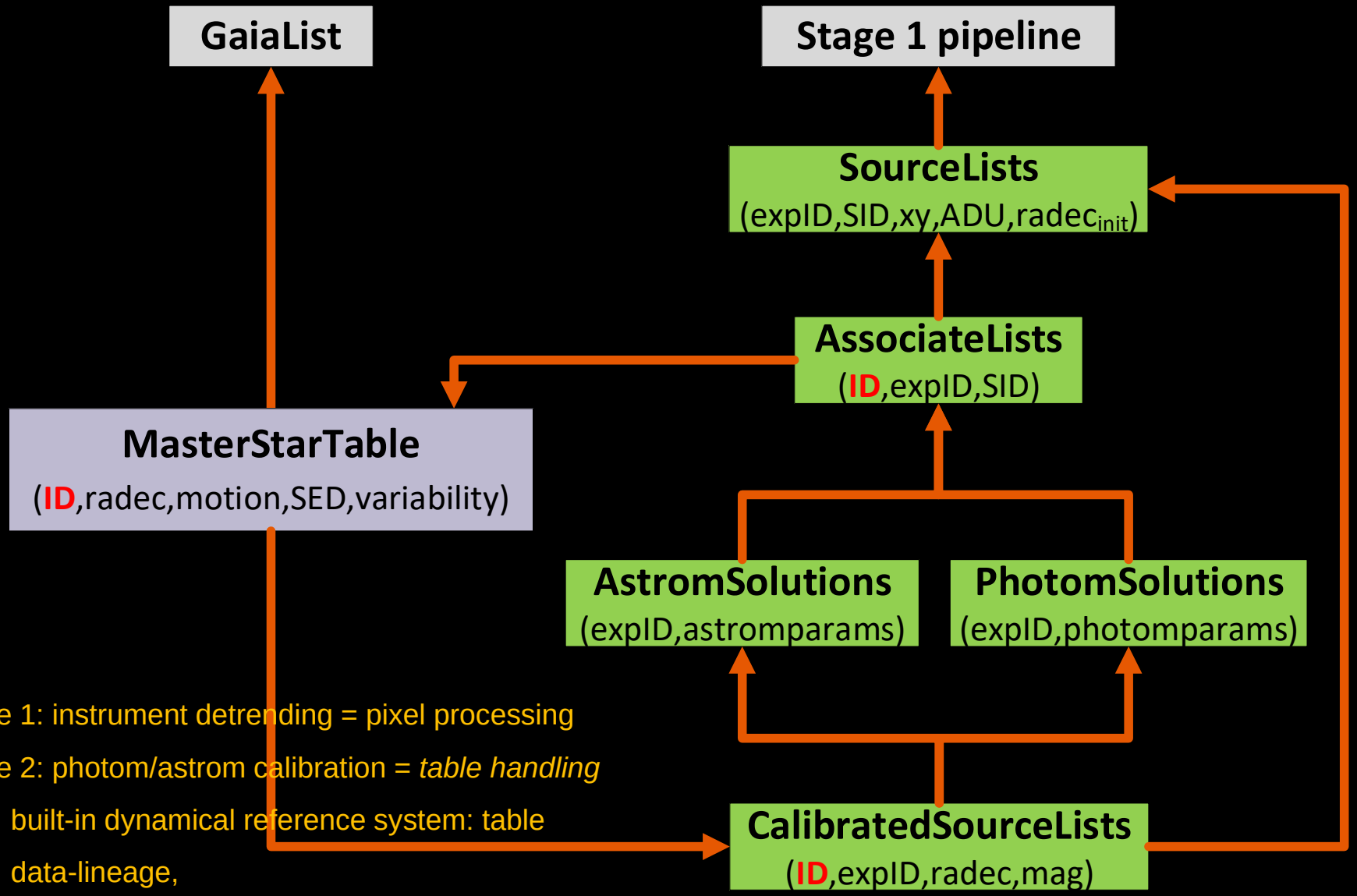






Euclidization
Changing reference systems
Astrometry- photometry

Target diagram (++ dependencies) for OU-EXT – Euclid external data - stage 2- dynamic Euclidization



Stage 1: instrument detrending = pixel processing

Stage 2: photom/astrom calibration = *table handling*

- built-in dynamical reference system: table data-lineage,
- QC, re-processing

Beyond Big Data

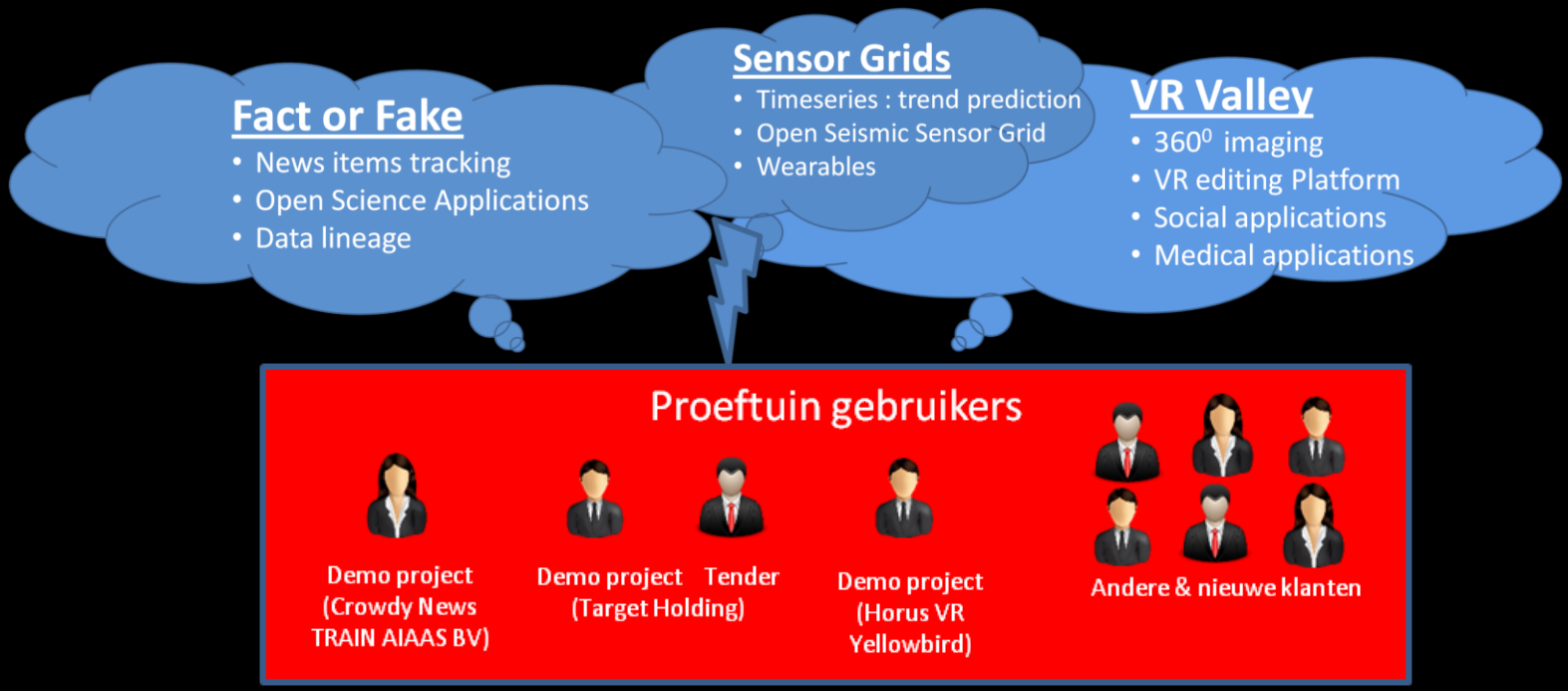
- QC and re-processing – Kids Euclid **FAIR**
- OU EXT > Billion – dynamic tables

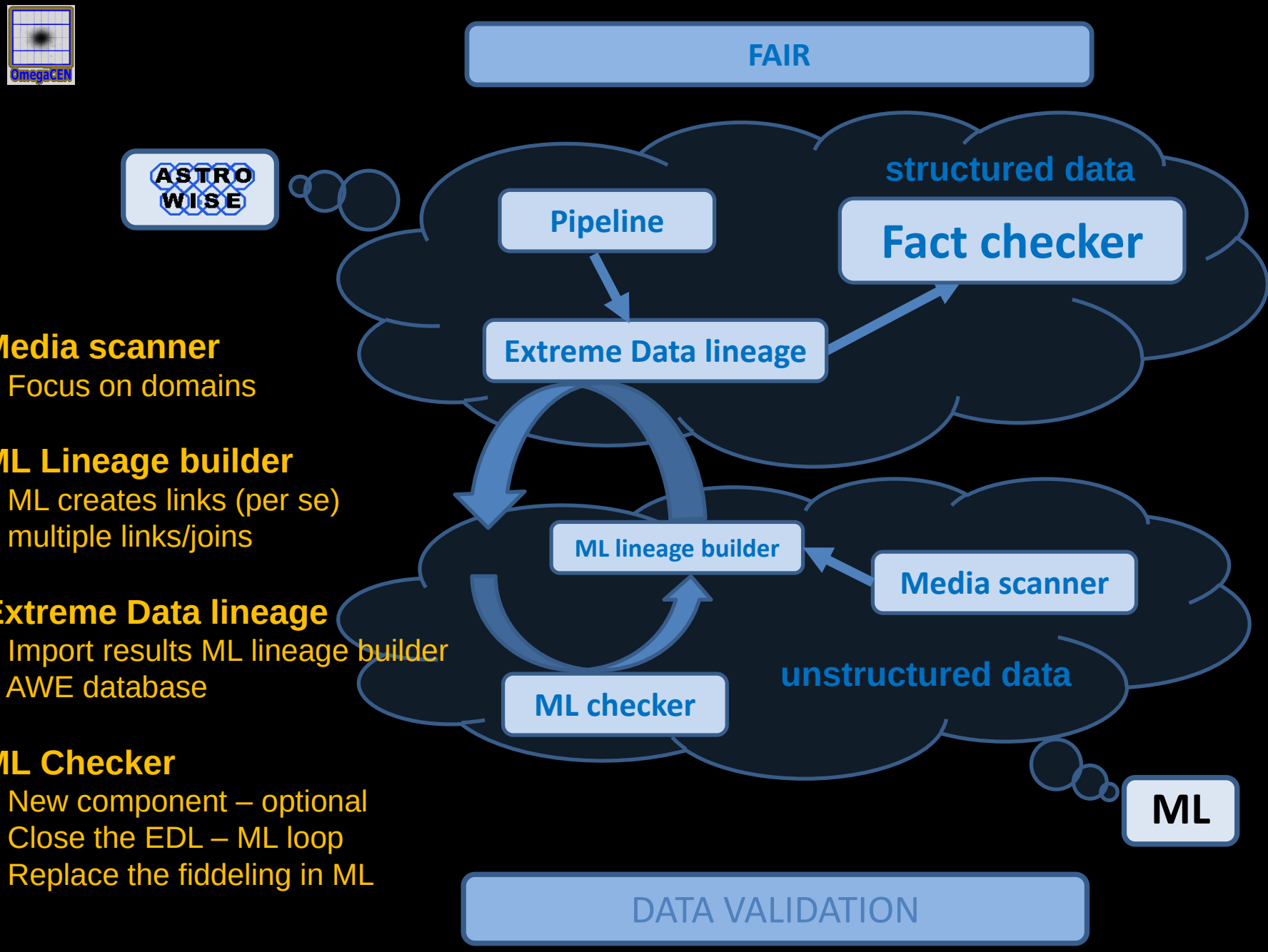
All techniques go back to the source

Scientists and journalists- > Fact and Fakes

Structured data and unstructured data

TARGET Fieldlab





Media scanner
Focus on domains

ML Lineage builder
ML creates links (per se)
multiple links/joins

Extreme Data lineage
Import results ML lineage builder
AWE database

ML Checker
New component – optional
Close the EDL – ML loop
Replace the fiddeling in ML

conclusions

Next level is all about Data validation

- check ML
- QC
- systematics in data sets
- OU-ext dynamic Euclidization
- unstructured data: ML + lineage

Almost all about going back to the source

Facts and Fakes